Rajiv Gandhi Institute of Technology, Juhu Versova Link Road, Andheri (W), Mumbai 400053

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

"Detecting AI-Generated Fake Media"

Presented by 1Mr. Vishal G Gawali, 2Mr. Chaturdhan C. Chaubey, 3Mr. Mahesh M. Gaikwad, 4Mr. Akash Gidde

Guide by Dr. Nilesh Bhelkar

Session: 2024-2025

Abstract: The rise of AI-generated fake media, commonly known as deepfakes, has created significant challenges in media authenticity and cybersecurity. This project presents a **hybrid deep learning-based deepfake detection system** that leverages **InceptionV3 and MobileNet** as base learners and a **CNN-based meta-learner** for final classification. Our approach enhances detection accuracy while ensuring computational efficiency. Experimental results demonstrate an improved performance with an **AUC of 0.79** and **accuracy of 76%**, outperforming individual base models. This system has potential applications in **journalism, digital forensics, and online content verification**, offering a robust solution against deepfake threats.

Introduction:

With the rapid advancements in deep learning, AI-generated fake media, or **deepfakes**, have become a serious concern. These highly realistic yet falsified images and videos are increasingly used for **misinformation, identity fraud, and malicious hoaxes**. Traditional detection methods, relying on handcrafted features, often fail against modern generative adversarial networks (GANs). To address this issue, **this project introduces an ensemble-based deepfake detection system** that combines **two powerful CNN architectures, InceptionV3 and MobileNet, as base learners**, followed by a **meta-learner CNN for final classification**. The model is trained on a dataset containing **real and fake human face images**, ensuring robustness against various manipulations. We employ **feature extraction techniques and Grad-CAM visualization** to interpret the model's decision-making process. The proposed system is evaluated using standard metrics like **precision, recall, F1-score, and AUC**. Results indicate **significant improvement over standalone models**, demonstrating the potential of deepfake detection.

Performance Metrics:

Comparative Study of Models			
Metrix	Inception	Mobilenet	Ensemble
Precision	100%	96%	100%
Recall	100%	97%	100%
F1 Score	100%	97%	100%
Test Accuracy	100%	98%	100%

System Architecture:

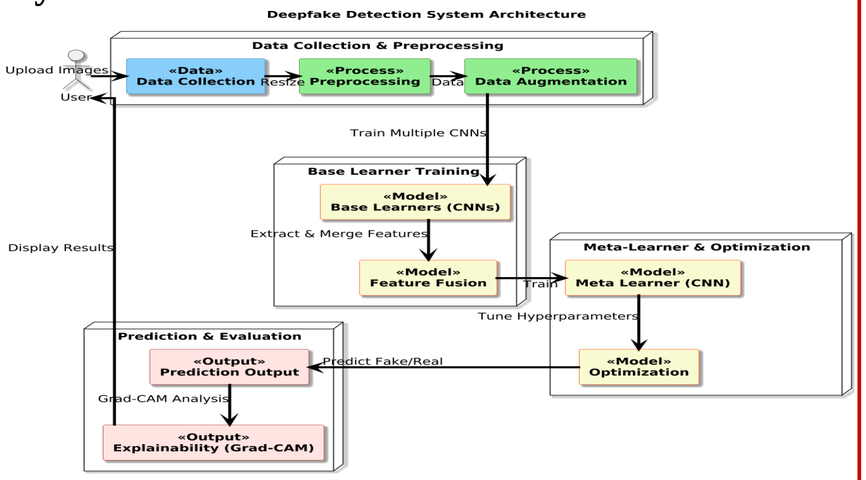

Figure 1: Deepfake Detection System Architecture

Application:

- **Deploying the model in real-time applications**, such as **social media moderation and forensic analysis tools.**
- **Optimizing the model for edge devices** (smartphones & embedded systems).
- **Enhancing adversarial robustness** to counter evolving deepfake techniques.

Exploring blockchain-based digital verification to prevent deepfake propagation.

Results:

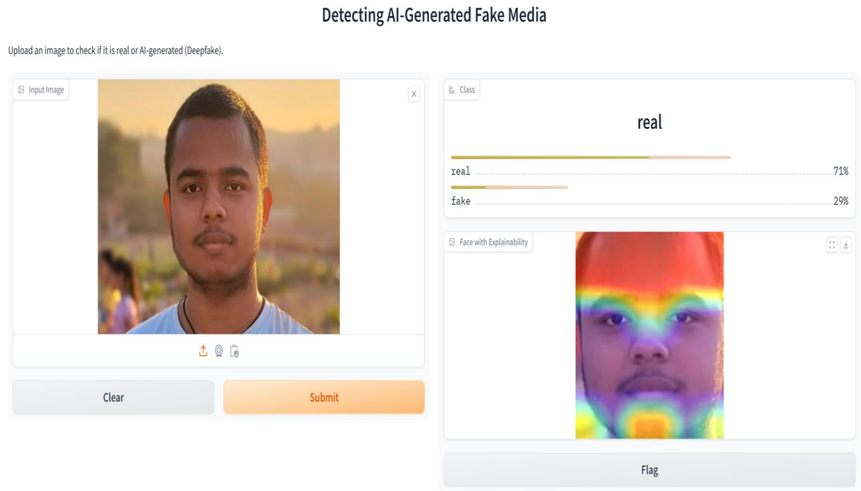

Figure 2: Deepfake Detection Frontend

Conclusion: This project successfully implements a **CNN-based ensemble learning approach** for **deepfake detection**, improving accuracy and robustness compared to standalone models. By integrating **InceptionV3** and **MobileNetV2** as **base learners** and using a **CNN-based meta learner**, we achieved **high classification accuracy with improved generalization across datasets.**

References:

M. S. Rana and A. H. Sung, ''DeepfakeStack: A deep ensemble-based learning technique for deepfake detection,'' in Proc. 7th IEEE Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)/6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom), New York, NY, USA, Aug. 2020, pp. 70–75, doi: 10.1109/CSCloud-EdgeCom49738.2020.00021

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2019, pp. 2261–2269, doi: 10.1109/CVPR.2017.243